

# Distributed Processing of Generalized Graph-Pattern Queries in SPARQL 1.1

Sairam Gurajada<sup>1</sup> and Martin Theobald<sup>2</sup>

<sup>1</sup> Max-Planck-Institute for Informatics, [gurajada@mpi-inf.mpg.de](mailto:gurajada@mpi-inf.mpg.de)

<sup>2</sup> University of Ulm, [martin.theobald@uni-ulm.de](mailto:martin.theobald@uni-ulm.de)

**Abstract.** We propose an efficient and scalable architecture for processing generalized graph-pattern queries as they are specified by the current W3C recommendation of the SPARQL 1.1 “Query Language” component. Specifically, the class of queries we consider consists of sets of SPARQL triple patterns with labeled property paths. From a relational perspective, this class resolves to conjunctive queries of *relational joins* with additional *graph-reachability predicates*. For the scalable, i.e., distributed, processing of this kind of queries over very large RDF collections, we develop a suitable partitioning and indexing scheme, which allows us to shard the RDF triples over an entire cluster of compute nodes and to process an incoming SPARQL query over all of the relevant graph partitions (and thus compute nodes) in parallel. Unlike most prior works in this field, we specifically aim at the *unified optimization and distributed processing* of queries consisting of both relational joins and graph-reachability predicates. All communication among the compute nodes is established via a proprietary, asynchronous communication protocol based on the Message Passing Interface.

## 1 Introduction

### 1.1 Background & Motivation

RDF [1] and SPARQL [2] are two standards recently recommended by the W3C for representing and querying linked data on the Web. RDF is a very versatile data format and hence found a wide adoption in many communities. Due to its broad applicability, it is frequently employed as a generic, albeit simple, knowledge representation format for large-scale information-extraction endeavors such as DBpedia, Freebase or YAGO, for capturing relationships in social networks, and many others. Consequently, much recent work has focused on the distributed, and hence scalable, processing of *graph-pattern queries* in SPARQL 1.0 via a variety of SQL [20,22] and NoSQL architectures [4,14]. Engines such as SHARD [19], SW-Store [15] (which is a Hadoop extension of the centralized RDF-3X [17] engine), EAGRE [27], Trinity.RDF [26], our own TriAD [11] engine, as well as commercial tools like Virtuoso [5] and Ontobroker [7] evaluate SPARQL queries either as series of relational joins [11,15,19,27] or via iterative graph exploration and message passing [26].

With its recent update, 1.1, SPARQL underwent a major revision. From a relational perspective, the introduction of *property paths* [3] likely constitutes the most remarkable change. Property paths allow for annotating pairs of query vertices by regular expressions in which properties and entire paths may be marked by a Kleene “+” or “\*”. SPARQL 1.1 thus introduces a notion of *generalized graph-pattern queries* in which property paths express transitive reachability constraints among sets of RDF entities that become bound to the query variables. That is, unlike in SPARQL 1.0, joins among triple patterns may be recursive and need to be evaluated either iteratively, or they involve a potentially costly materialization of the properties’ closures.

As opposed to the large variety of SPARQL 1.0 engines, the processing of property paths in SPARQL 1.1 so far has been investigated by only very few approaches [5,10,18] (of which only [5] is available). A particular challenge lies in the combined optimization of relational joins among the triple patterns (based on shared variables) with additional graph-reachability predicates (based on one or more properties marked by a “+” or “\*”). While state-of-art indexing techniques for graph-reachability predicates [9,21,25] are inherently limited to a centralized setting, we are currently aware of only one approach that specifically tackles distributed reachability queries for single-source, single-target queries [6]. For multi-source, multi-target reachability queries, as they frequently occur in SPARQL 1.1, we are aware of just two centralized approaches [9,23] that provide suitable indexing and processing strategies. Distributed graph engines, such as Berkeley’s GraphX [24] and Apache Giraph, on the other hand, allow for the scalable processing of graph queries over massive, partitioned data graphs. Both provide generic API’s for implementing various kinds of graph algorithms, including multi-source, multi-target reachability queries. However, they do not support the kinds of indexing techniques known from the centralized approaches, and they are not directly amenable to the declarative way of querying RDF data as it is suggested by the SPARQL standard. Processing property paths with transitive reachability constraints under these frameworks requires an iterative form of graph traversal, which may result in as many iterations (and hence communication rounds) as the diameter of the graph in the worst case.

Summarizing this motivating section, we are not aware of any previous approach that achieves a true scale-out in processing generalized graph-pattern queries for SPARQL 1.1 under both *strong* (thus reducing the query time by increasing the number of processors) and *weak scaling* (thus achieving predictable query times when increasing both the data size and the number of processors). In this paper, we present a distributed architecture that takes advantage of indexing techniques for graph-reachability predicates known from centralized approaches, while retaining the declarative style of querying massive RDF data graphs for SPARQL 1.1 with property paths.

## 1.2 Contributions

We summarize the contributions of our work as follows.

- We present an *efficient* and *scalable query engine* for the core of the SPARQL 1.1. Specifically, the fragment of SPARQL we consider resolves to sets of triple

patterns with labeled property paths, which allows for formulating generalized graph pattern queries as conjunctive queries of relational joins with additional graph-reachability predicates (including one or more properties marked by a Kleene “+” or “\*”).

- We provide a *unified indexing scheme*, *cost model* and *query optimization framework* to seamlessly integrate graph-reachability predicates into the relational query processor of our TriAD [11] engine. TriAD\* employs a strictly fixed, asynchronous message-passing protocol to evaluate a SPARQL 1.1 query among all of the compute nodes in parallel. Our protocol requires exactly one round of communication per graph-reachability predicate and thus avoids a costly, iterative form of communication.
- Our approach is the first to report an actual *horizontal scale-out* in processing SPARQL 1.1 queries with property paths over a number of large RDF collections. We present a detailed experimental evaluation of our approach under both strong and weak scaling, and in comparison to the Virtuoso native RDF store.

## 2 RDF Data & SPARQL Query Model

We next formally define our data and query model. Our model is based on a graph representation of RDF data and supports the core syntax of SPARQL 1.1 to express queries over an RDF data graph as graph-pattern queries. Specifically, we focus on the **WHERE** clause of a SPARQL 1.1 query, in which triple patterns (including triples annotated with property paths) are connected (i.e., “joined”) via their shared variables. We currently omit other SPARQL 1.0/1.1 extensions such as **UNION**, **FILTER**, aggregations, subqueries and negation.

### 2.1 RDF Data Model

We follow the general W3C recommendation of the basic RDF vocabulary [1] (however ignoring RDF schema extensions), in which an RDF collection consists of a *set of triples* which are each of the form  $\langle \text{subject}, \text{property}, \text{object} \rangle$  (or  $\langle s, p, o \rangle$ , for short). A set of RDF triples can thus concisely be represented as a directed, labeled multi-graph, as it is defined next.

**Definition 1.** An **RDF data graph**  $G_D(V_D, E_D, Cons, \phi_D)$  is a directed, labeled multi-graph where  $V_D$  is the set of data vertices,  $E_D$  is the set of directed edges connecting the vertices in  $V_D$ ,  $Cons$  is the set of vertex and edge labels, and  $\phi_D$  is a labeling function with  $\phi_D : V_D \cup E_D \rightarrow Cons$  s.t.  $\forall v_i, v_j \in V_D, v_i \neq v_j$ , it holds that  $\phi_D(v_i) \neq \phi_D(v_j)$ .

**Partitioning RDF Data Graphs.** A potentially scalable approach for managing very large RDF collections is to follow a distributed indexing scheme. In our approach, an RDF data graph is partitioned into multiple graph partitions, such that one or more of these graph partitions can be managed locally by a compute node. More formally, an RDF data graph  $G_D$  is partitioned into  $k$  vertex- and edge-disjoint graphs  $\mathcal{G} = \{G_1, G_2, \dots, G_k\}$ , where each  $G_i(V_i, E_i, Cons, \phi_D)$  is a *vertex-induced subgraph* of  $G_D$ . That is, it holds that  $\bigcup_{i=1..k} V_i = V_D$  and

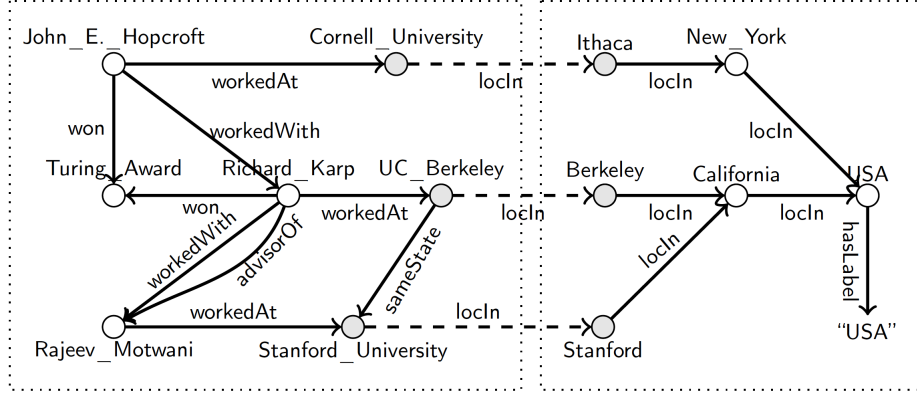


Fig. 1. A partitioned RDF data graph

$V_i \cap V_j = \emptyset$  for all  $i \neq j$ . Figure 1 shows an example RDF data graph that is partitioned into the two graph partitions  $G_1$ ,  $G_2$  which are located at slaves 1 and 2, respectively.

## 2.2 SPARQL Query Model

SPARQL 1.0 [2] is a widely adopted W3C language recommendation for querying RDF data. A typical SPARQL 1.0 query comprises a *conjunction of triple patterns* of the form  $\langle s, p, o \rangle$ , where each of the  $s, p, o$  components may denote either a constant or a variable. For instance, the query “Find all professors who won a Turing Award and worked in a US university” can be specified in SPARQL 1.0 as follows.

```
SELECT ?person WHERE {
  ?person won Turing_Award. ?person workedAt ?univ.
  ?univ locIn ?city. ?city locIn ?state.
  ?state locIn ?country. ?country hasLabel "USA" }
```

**Property Paths.** With the latest revision, SPARQL 1.1, *property paths* [3] were introduced to SPARQL 1.0. A property path specifies how an entity is (possibly transitively) connected to another entity. The above query can thus concisely be rewritten by using a property path as follows.

```
SELECT ?person WHERE {
  ?person won Turing_Award.
  ?person workedAt/locIn */hasLabel "USA" }
```

Here, the property path `workedAt/locIn */hasLabel` expresses a transitive connectivity via the `locIn` property. That is, a path connecting any vertex that becomes bound to the variable `?person` with the vertex with label “USA” must involve zero or more consecutive properties with the label `locIn`. Property paths thus help the user in two ways: (1) in simplifying the query representation and (2) in formulating queries with partial knowledge of the underlying schema.

**SwPP Queries.** We focus on a subset of queries expressible in SPARQL 1.1, in the following called “SwPP”, consisting of *conjunctions of triple patterns with property paths*. Specifically, a triple pattern of an SwPP query again is of the

form  $\langle s, p, o \rangle$ , where each of the  $s, o$  components may refer to either a constant in  $Cons$  or to a query variable from a distinct set of variables  $Vars$ . Moreover,  $p$  may refer to a **path** expression whose grammar we define below.

<b>path</b> := <b>path/path</b>	(concatenation of paths)	(1)
:= <b>URI</b>	(single property)	(2)
:= <b>^URI</b>	(inverse property direction)	(3)
:= <b>URI?</b>	(zero or one property)	(4)
:= <b>URI*</b>	(zero or more properties)	(5)
:= <b>URI+</b>	(one or more properties)	(6)

Formally, an SwPP query thus again forms a directed, labeled multi-graph, as it is defined next.

**Definition 2.** A **SPARQL query graph**  $G_Q (V_Q, E_Q, Cons, Vars, \phi_Q, \psi_Q)$  is a directed, labeled multi-graph where  $V_Q$  is the set of query vertices,  $E_Q$  is the set of edges connecting the vertices in  $V_Q$ ,  $Cons$  is the set of vertex and edge labels,  $Vars$  is a set of query variables,  $\phi_Q$  is a vertex-labeling function with  $\phi_Q : V_Q \rightarrow Cons \cup Vars$ , and  $\psi_Q$  is an edge-labeling function with  $\psi_Q : E_Q \rightarrow \{ \wedge \} \mid Cons \mid \{ *, +, ? \}$ .

We hereby adopt a simpler definition for property paths than the full syntax proposed by the W3C [3]. However, by rewriting an entire path expression (denoted as “**path**” in the above grammar) into a sequence of join conditions, each with a property that denotes a single URI (referred to as “**URI**”) with an optional Kleene “**\***” or “**+**”, we allow a more general syntax for property paths rather than just a single transitive property. The above grammar in particular allows concatenations of properties into paths of arbitrary length, as long as these can be rewritten into a conjunction of triple patterns and thus conform to Definition 2. In our implementation, the distinction between “**+**”, “**\***”, “**?**” and “**^**” is very simple. For “**+**”, we merely disallow an equality between a source and a target vertex; while for “**?**”, we restrict the maximum path length to 1. The inversion “**^**” simply swaps the source and target vertices.

### 3 Architecture

The architecture of our SPARQL 1.1 engine (in the following coined TriAD\*) is based on TriAD [11,12], which we originally developed for processing conjunctions of triple patterns in SPARQL 1.0. The design of TriAD in principle follows a classical master-slave architecture at indexing time, but allows for a direct, asynchronous communication among the slaves at query-processing time.

**Indexing.** At indexing time, the master node is responsible for the encoding of incoming RDF triples and for the distribution of these triples among the slave nodes. The slaves then construct their local index structures (each over a distinct graph partition) and send their local index statistics back to the master node.

**Query Processing.** At query time, the master node facilitates these precomputed index statistics to compile a global query plan that is then distributed to and processed at all slaves in parallel. While processing such a query plan, the slaves directly exchange their intermediate query results via a proprietary, asynchronous communication protocol based on the Message Passing Interface [8].

To facilitate the processing of SPARQL 1.1 queries with property paths, TriAD\* augments the architecture of TriAD in two decisive ways:

- **Triple & Reachability Indexes:** We create two kinds of index structures—one that is optimized for processing joins and another one that is optimized for graph-reachability queries among the slaves (see Section 4).
- **Unified Query Optimization:** We devise a unified cost model for optimizing both relational joins and graph-reachability predicates (see Section 5).

## 4 Index Organization

We use a hashing-based sharding scheme to distribute RDF triples across the slaves [11,15,19]. Each RDF triple in a collection is parsed at the master node and encoded into an integer format via a dictionary to obtain compact identifiers for each of the  $s$ ,  $p$  and  $o$  components. Every such encoded  $\langle s, p, o \rangle$  triple is then distributed to (at most) two slaves  $i, j$  by choosing  $i = (s \bmod k)$  and  $j = (o \bmod k)$  as sharding conditions, respectively. Although we use a hash function for sharding, we can easily adopt different partitioning schemes, for example, by exchanging  $s, o$  in the hashes with the graph partitions to which  $s, o$  are assigned by a locality-preserving graph-partitioning framework such as METIS [16].

### 4.1 Triple Indexes

Following TriAD’s general indexing strategy [11], each slave creates six permutations of in-memory triple indexes for the efficient processing of relational joins among triple patterns based on their subjects or objects. Depending on whether an incoming triple was hashed to slave  $i$  via its subject or object, these six permutations are arranged into two groups: (i) the three *subject-key* indexes (SPO, SOP, PSO) and (ii) the three *object-key* indexes (OSP, OPS, POS). That is, a sharded triple, which was hashed onto slave  $i$  via its subject field, is indexed in three permutations by using the subject as key. Likewise, a triple hashed onto slave  $i$  via its object field is indexed three times by using the object as key. The two PSO and POS permutations are employed for index scans of triple patterns with a given property and an optional subject or object as key.

**Sorting & Scans.** The six vectors storing the SPO permutations at each slave are each sorted with respect to their permutation order. For example, the OPS vector is sorted primarily by the object field, followed by the predicate field, and finally by the subject field. Thus, when using a given object as search key, we first determine the entry point in the OPS vector via a binary search and then start scanning all triples under this search key sequentially.

### 4.2 Indexes for Property Paths

In addition to the triple indexes, each slave constructs a separate reachability index to facilitate the efficient processing of property paths in SPARQL 1.1.

One such reachability index is built for each unique property label  $p$  in the RDF collection, similar to the approach described in [10]. The construction of our reachability index for a property  $p$  over a partitioned RDF graph is based on our recent results from [13], which we briefly summarize below.

Let  $\mathcal{G}^p = \{G_1^p, G_2^p, \dots, G_k^p\}$  be the  $p$ -induced subgraphs of an RDF data graph  $G_D$ . That is, each  $G_i^p$  is a subgraph of a previously determined graph partition  $G_i$  (e.g., by using METIS [16] to partition the original multi-labeled data graph  $G_D$ ) that comprises of only  $p$ -labeled edges. For each such subgraph  $G_i^p$ , we first identify the set of vertices  $I_i^p$  (coined “*in-boundaries*”) and  $O_i^p$  (coined “*out-boundaries*”) that lie on the cut  $C$  that is given by  $\mathcal{G}^p$ . Next, for each partition  $G_i^p$ , we precompute all reachable pairs of vertices from in-boundaries  $I_i^p$  to out-boundaries  $O_i^p$ . We then construct two graph-based index structures, called *boundary graph* and *compound graph* (described in detail in [13]), as follows. Each set of reachable pairs for a partition  $G_i^p$  is represented as a bipartite graph with edges from in-boundaries to out-boundaries that are transitively reachable. Boundary nodes which are both in- and out-boundaries are omitted from this step. The bipartite graphs for each partition  $G_i^p$  are then communicated to (i.e., replicated across) all slaves  $j$  (for all  $j \neq i$ ). At each slave  $i$ , a *boundary graph*  $B_i^p$  is constructed by merging the received bipartite graphs from slaves  $j$  (for all  $j \neq i$ ) with the cut  $C$ . Finally, we construct a compound graph  $C_i^p$  by merging the boundary graph  $B_i^p$  also with the local subgraph  $G_i^p$ .

**Reachability Indexes.** To reduce the sizes of the precomputed boundary graphs  $B_i^p$ , on the one hand, and to further accelerate graph-reachability queries over the compound graphs  $C_i^p$ , on the other hand, we apply the following optimizations (also described in detail in [13]). First, we determine equivalence sets over in- and out-boundaries  $I_i^p$  and  $O_i^p$ . The in- and out-boundaries are then redefined with respect to new virtual vertices, each representing an equivalence set. That is,  $I_i^p$  and  $O_i^p$  comprises of all *in-virtual* and *out-virtual* vertices, respectively. The bipartite graphs are then computed over the redefined  $I_i^p$  and  $O_i^p$  sets. This technique helps to reduce the size of the boundary graphs  $B_i^p$ . The compressed boundary graphs are then used to construct the compound graphs  $C_i^p$  at each slave  $i$  in the same manner as before. Second, the obtained compound graphs are each condensed into a *directed acyclic graph* (DAG), in which each vertex represents a strongly connected component (SCC) of the compound graph. We finally store these DAGs as in-memory adjacency lists (serialized into another pair of PSO and POS vectors) and perform a local DFS traversal to answer a graph-reachability query at each slave.

### 4.3 Index Statistics

To optimize SwPP queries consisting of both relational joins among triple patterns and of triple patterns with property paths, we rely on a cost-based plan generator which is part of TriAD’s architecture. For this, we collect various statistics over the RDF data for both the basic triple patterns and triple patterns with property paths.

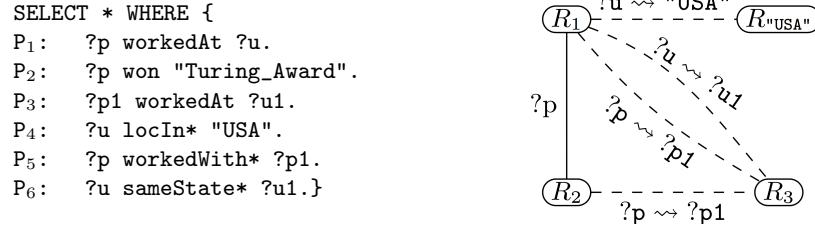
**Statistics for Triple Patterns.** As in [11], our statistics for basic triple patterns include (i) cardinalities  $Card(R_i)$  of relations  $R_i$  induced by individual

*subject*, *property* and *object* keys and (ii) of relations induced by *subject-object*, *property-subject* and *property-object* pairs. In addition, we compute the join selectivities  $Sel(p_i, p_j)$  of all *pairs of properties*  $p_i, p_j$  to estimate the cardinality of a join among two triple patterns.

**Statistics for Property Paths.** In order to plug triple patterns with property paths into our optimizer, we need to also estimate the selectivity of a property path. Precomputing these selectivities for every possible property path that may occur in a query is clearly intractable. We thus follow a simple sampling-based approach. For each individual property  $p$ , we determine the *reachability selectivity*  $Sel(p)$  as the fraction of randomly sampled source and target vertices  $(s, t)$ , for which  $s \rightsquigarrow t$  holds with respect to the subgraph of  $G_D$  that is induced by  $p$ .

## 5 Query Optimization & Distributed Processing

We first translate SwPP queries in SPARQL 1.1 into a graphical representation, which then forms the basis for query optimization. These query graphs are generated by introducing a vertex for each triple pattern in the query, while the edges that connect two such vertices represent equi-joins. These equi-joins are based on the variables (i.e., either the subjects or objects) that are shared by two such triple patterns. Edges for equi-joins are labeled with the shared variables. In addition to SPARQL 1.0 queries, SwPP queries contain triple patterns with property paths. Following [10], we represent a property path by a distinguished edge among two such query vertices, whose labels denote the reachability predicates among the subjects or objects in the respective vertices' triple patterns. In case the subject or object of a connected triple pattern is either a constant or an unbound variable (i.e., the variable is not present in the other triple patterns), we create a new query vertex for the same and add an edge between the respective query vertices.



**Fig. 2.** Example SwPP query and its graph representation

Figure 2 shows an example SwPP query and its corresponding graph representation. Here, patterns  $P_1, P_2, P_3$  are basic triple patterns whose property each consists of a single URI. These are represented as vertices  $R_1, R_2, R_3$ , respectively, in the query graph. Since, the property path of  $P_4$  points to only the constant "USA", a separate vertex  $R_{\text{"USA"}}$  representing a relation that consists of just a singleton tuple is added to the query graph. The equi-join on the shared variable  $?p$  is represented by the continuous line between  $R_1$  and  $R_2$ . A reachability edge, denoted by a dashed line for each property path, is added between



the respective subjects' and objects' query vertices. This is the case between  $R_1$  and  $R_{\text{"USA"}}$  for the property path of  $P_4$ , between  $R_1$  and  $R_3$  for the property paths of  $P_5$  and  $P_6$ , and between  $R_2$  and  $R_3$  for  $P_5$ .

### 5.1 Query Optimization

Once the query is translated into its graph representation, classical join-order-enumeration techniques [11,17] can be employed to find a cost-efficient execution plan. We extend TriAD's optimizer to handle SwPP queries by adding a new operator—*Distributed Reachability Join* (DRJ)—and respective cost estimator for property paths. Next, we briefly discuss these operators, which is followed by a discussion of the cost estimation and join-order enumeration.

**Physical Query Operators.** TriAD employs three physical operators—coined Distributed Index Scan (DIS), Distributed Merge Join (DMJ) and Distributed Hash Join (DHJ)—for processing index scans and equi-joins among triple patterns in SPARQL 1.0. Each of these operators works over the sharded partitions of the triple indexes described in Section 4.1 in parallel. In short, the DIS operators, which only occur at the leaves of the query plan, each build a relation by invoking a parallel scan over the respective SPO permutation index that was selected by the optimizer. The DMJ and DHJ operators each take two sharded relations plus the join keys (i.e., the shared variables) as input and perform a hash- or merge-join, respectively, to generate a new intermediate relation.

**Distributed Reachability Join (DRJ).** Analogously, we define a new DRJ operator to process triple patterns with property paths in TriAD\*. This enhanced join operator takes two sharded relations  $R_i, R_j$  as input and returns as output the subset of tuples in the cross-product  $R_i \times R_j$ , for which all of the attached *join conditions*  $\mathcal{C}$  hold:

- for each shared variable  $?x$  in  $\mathcal{C}$ , a pair of tuples in  $R_i$  and  $R_j$  must have equal values for  $?x$ ; and
- for each reachability predicate  $?x \rightsquigarrow ?y$  in  $\mathcal{C}$ , a vertex  $s$  that becomes bound to  $?x$  by a tuple in  $R_i$  must be reachable to a vertex  $t$  that becomes bound to  $?y$  by a tuple in  $R_j$ .

The evaluation of the DRJ operator is backed by the index structures for property paths described in Section 4.2. That is, for sets of source and target vertices that become bound to the variables of a reachability predicate  $?x \rightsquigarrow ?y$ , the DRJ operator performs a distributed set-reachability operation using the pre-computed graph indexes over all partitions in parallel.

**Join-Order Optimization.** The DP table of the optimizer is initialized with the cost estimates for the DIS operations of each query vertex  $R_i$ . The scan costs for  $R_i$  depend on whether the constants in the triple pattern match a respective SPO permutation index  $idx$ . For instance, if the subject and predicate are constants and the object is a variable, choosing an SPO or PSO permutation costs much less compared to any of the remaining permutations.

In the query graph, we introduce two special kinds of query vertices, namely one for property paths with a constant subject or object, and one for property

paths with (at least one) unbound variable. Scanning a singleton tuple as input has a unit cost of 1, while scanning a relation constructed for an unbound variable corresponds to the number of triples in the subgraph  $G^p = (V^p, E^p)$  of  $G_D$  that is induced by property  $p$ . In the latter case, we thus set the cardinality  $Card(R_i)$  of a unary relation  $R_i$  that is constructed from all vertices in  $G^p$  to  $|V^p|$ . As an example, consider the property path `?x locIn* ?y`, and let variable `?y` be unbound (i.e., not occurring as a shared variable in any other triple pattern). Then the number of unique bindings for `?y` is the number of vertices in the induced subgraph consisting only of `locIn` edges. Equation 7 summarizes the cost estimates we obtain for a DIS operator with respect to the precomputed cardinalities  $Card(R_i)$  and available SPO permutations.

$$Cost(R_i) \propto \begin{cases} 1 & \text{if } R_i \text{ is a singleton tuple;} \\ Card(R_i^{idx})/k & \text{if } R_i \text{ is sharded across } k \text{ slaves} \\ & \text{and matches SPO index } idx; \\ \infty & \text{otherwise.} \end{cases} \quad (7)$$

Once the DP table is initialized with the costs estimates for the DIS operators, we continue to build the join tree in a bottom-up manner. At each DP step, we merge two branches  $Q^{left}$ ,  $Q^{right}$  into a combined plan  $Q$  by a join operator  $op$  together with a set of join conditions  $\mathcal{C}$ . If there is at least one reachability edge between two relations  $R_i$ ,  $R_j$  that connect  $Q^{left}$  and  $Q^{right}$ , a DRJ operator is employed. Note that each DRJ operator may invoke multiple distributed set-reachability queries depending on the number of edges that connect  $R_i$  and  $R_j$ . That is, for the case  $Q^{left} := \bowtie^{DMJ} (R_1, R_2)$  and  $Q^{right} := R_3$  shown in Figure 2, we need the DRJ operator to consider two reachability predicates `?p workedWith* ?p1` and `?u sameState* ?u1`. Choosing the correct order for executing multiple reachability predicates thus is a sub-goal of overall optimization. Assuming independence among the join conditions  $\mathcal{C}$ , we plug in our precomputed index statistics as follows.

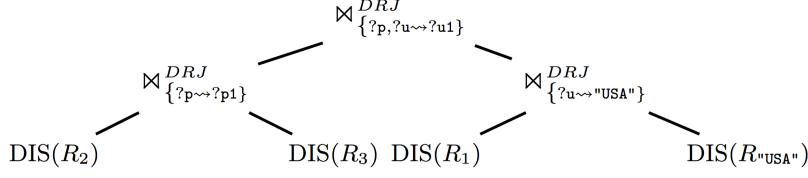
$$Cost(Q^{left} \bowtie_{\mathcal{C}}^{op} Q^{right}) \propto \sum_{C_i \in \mathcal{C}} Card(Q_i^{left}) \cdot Card(Q_i^{right}) \cdot Sel(C_i) \quad (8)$$

While processing the conditions  $C_i \in \mathcal{C}$ , we also iteratively estimate the cardinality  $Card(Q_i)$  of a subquery  $Q_i$  of Equation 8 as defined next.

$$Card(Q_i) := \begin{cases} Card(Q_1) & \text{if } i = 1 \\ \prod_{j=1}^{i-1} Card(Q_j) \cdot Sel(C_j) & \text{if } i > 1 \end{cases} \quad (9)$$

Thus, if  $C_j$  is a graph-reachability predicate,  $Sel(C_j)$  denotes the reachability selectivity  $Sel(p)$  of the property  $p$  that is associated with  $C_j$ . If  $C_j$  refers to an equi-join,  $Sel(C_i)$  denotes the precomputed join selectivity  $Sel(p_i, p_j)$  for the pair of properties associated with the two triple patterns of the equi-join. The combined cost for a (sub-)query  $Q$  then is defined recursively.

$$Cost(Q) = \begin{cases} \max(Cost(Q^{left}), Cost(Q^{right})) \\ + Cost(Q^{left} \bowtie_{\mathcal{C}}^{op} Q^{right}) \\ + Cost(Q^{left} \stackrel{op}{=} Q^{right}) \end{cases} \quad (10)$$



**Fig. 3.** Example plan for the query of Figure 2

Here,  $Cost(Q^{left} \bowtie_C^{op} Q^{right})$  denotes the cost of processing the join operator  $op \in \{DMJ, DHJ, DRJ\}$  with  $Q^{left}$  and  $Q^{right}$  as operands and join conditions  $C$  (Eq. 8). Likewise,  $Cost(Q^{left} \rightleftharpoons^{op} Q^{right})$  accounts for the shipping costs that incur when the resharding of intermediate relations is required. The shipping cost is proportional to the size and width of  $Q^{left}$  and  $Q^{right}$ , respectively. Using  $\max(\cdot, \cdot)$  as cost aggregation finally accounts for the parallel execution of the two branches [11]. Figure 3 shows an example query plan for the query of Figure 2.

## 5.2 Distributed Query Processing

We embed the new DRJ operator into TriAD's multi-threaded and asynchronous processing framework to support the distributed execution SwPP queries. The principal processing flow and communication protocol [11] remain unchanged and merely require an additional initialization of the source and target vertices for the distributed set-reachability queries, which are now triggered by the DRJ operators at their respective positions in the query plan.

**1. Scanning Base Relations.** The leaves of the operator tree always represent distributed index scans (DIS). Each slave scans its local SPO permutation index and selects tuples according to the constants associated with the DIS operator. Due to the layout of our SPO indexes, this merely requires initializing an iterator at the first tuple in a permutation list that matches the constants. For a DRJ operator with a reachability predicate, whose source or target is a single constant, a singleton relation is created directly from that constant. If the DRJ operator has a reachability predicate with an unbound variable via a property  $p$ , a (sharded) unary relation with the local vertices of  $V^p$  is created.

**2. Query-Time Sharding.** During the execution of the query plan, resharding of intermediate relations may be required to ensure the proper execution of joins (DMJ, DHJ) and set-reachability (DRJ) operations. With six SPO permutations, each DMJ operator requires sharding of at most one of its base relations at query time, while the DHJ operator requires sharding of at least one of its intermediate relations, depending on the locality of the tuples with respect to the join key. Sharding for the DRJ operator depends on the locality of the join keys based on shared variables (if present) and the locality of the vertices that become bound to the source and target variables of the reachability predicates. Thus, resharding may be required for both input relations of a DRJ operator.

**3. Parallel Execution of Operators.** In addition to the concurrent execution of the operators across the slaves, each slave also locally pursues the execution of the query plan in a multi-threaded fashion. Starting from the leaves

of the query plan, all operators are locally executed in one separate thread for each *execution path* (EP) (i.e., for each distinct leaf-to-root path) in the query plan. Since slaves may take different amounts of time to execute an operator over their local partition of the index, an *asynchronous* exchange of messages for resharding the partial relations at query time makes this step more efficient than a synchronous protocol. As soon as all the shards for the two input relations of a join operator are in place, the threads of the two EPs at each slave are merged into one, and the next join operations can be invoked locally. For a DRJ operator with a graph-reachability predicate, whose source or target variables become bound to constants due to a shared variable, the respective source and target sets for the distributed set-reachability query are initialized from those constants. These are then resharded to the slaves that hold the graph partitions containing the source and target vertices.

## 6 Evaluation

TriAD\* is entirely implemented in C++. We used GCC-4.7.3 with `-O3` optimization and MPICH2-1.4.1 and Boost-1.55 as external libraries. We ran all of the following experiments on a compute cluster with up to 11 nodes, out of which 1 was dedicated as the master node. Each node runs Debian 7.5, has 48GB of RAM and an Intel E5530@2.40GHz quad core CPU with HT enabled.

**Datasets.** We used three large-scale, both real-world and synthetic, RDF datasets for our evaluation: (i) LUBM-500M<sup>3</sup> (scaled to 500 million triples) is generated using UBA 1.7 in N3 format, (ii) Freebase-500M (with 500 million triples) refers to a subset of a recent Freebase snapshot<sup>4</sup> and (iii) a recent snapshot of DBpedia<sup>5</sup> (with 417,445,957 triples).

**Queries.** We manually designed three queries for each dataset (L1–L3 for LUBM, F1–F3 for Freebase, D1–D3 for DBpedia) to capture a mixture of reachability queries and relational joins. All queries are listed in our Appendix.

### 6.1 Distributed SwPP Queries

We first discuss the distributed processing of SwPP queries for the fixed snapshot of the three datasets described above. For TriAD\*, we used 5 slaves for this setting (plus 1 master node). As competitor, we used the Virtuoso 7.1.0 native RDF store, which is the only available RDF store we are aware of that supports full property-path processing. We remark that the open-source edition of Virtuoso 7.1.0 does not support distribution. We thus compare against a centralized installation of Virtuoso on one of our compute nodes.

**A. LUBM-500M.** The results for processing SwPP queries (L1, L2, L3) are shown in Table 1(a). L1 resembles a single, non-selective reachability join. Processing L1 thus involves an index scan for two input relations and a respective evaluation of the reachability join. We can observe that the centralized version of TriAD\* performs better than Virtuoso in a *cold* cache and comparable to

<sup>3</sup> <http://swat.cse.lehigh.edu/projects/lubm/>

<sup>4</sup> <https://developers.google.com/freebase/data>

<sup>5</sup> <http://downloads.dbpedia.org/2015-04/core/>

(a) LUBM-500 ( <i>query times in seconds</i> )					
	#Slaves	L1	L2	L3	Geo.-Mean
TriAD*	1	6.437	0.331	42.681	4.497
TriAD*	5	<b>1.250</b>	<b>0.162</b>	<b>8.516</b>	<b>1.199</b>
Virtuoso (cold)	1	10.050	12.624	57.776	19.425
Virtuoso (warm)	1	4.963	5.452	56.603	11.527

(b) Freebase-500 ( <i>query times in seconds</i> )					
	#Slaves	F1	F2	F3	Geo.-Mean
TriAD*	1	1.084	1.568	0.677	1.048
TriAD*	5	<b>0.356</b>	0.642	<b>0.423</b>	0.459
Virtuoso (cold)	1	6.590	4.112	13.809	7.206
Virtuoso (warm)	1	1.196	<b>0.002</b>	5.601	<b>0.238</b>

(c) DBpedia ( <i>query times in seconds</i> )					
	#Slaves	D1	D2	D3	Geo.-Mean
TriAD*	1	24.822	0.713	29.407	8.044
TriAD*	5	<b>7.973</b>	<b>0.412</b>	<b>11.223</b>	<b>3.328</b>
Virtuoso (cold)	1	46.185	19.352	317.899	65.741
Virtuoso (warm)	1	27.820	2.395	302.753	27.222

**Table 1.** Performance evaluation of SwPP queries

Virtuoso in a *warm* cache setting. We however achieve a significant scale-out for L1 when we evaluate the query on a cluster of 5 slaves. Next, L2 is a selective query with two regular joins and a single reachability join. For this query, we can observe that TriAD\* achieves a better performance compared to Virtuoso in both the cold and warm cache settings. The non-selective query L3 contains two reachability joins in conjunction with two regular joins. Also here, TriAD\* continues to perform better than Virtuoso under both a cold and warm cache and further scales out very well in a distributed setting.

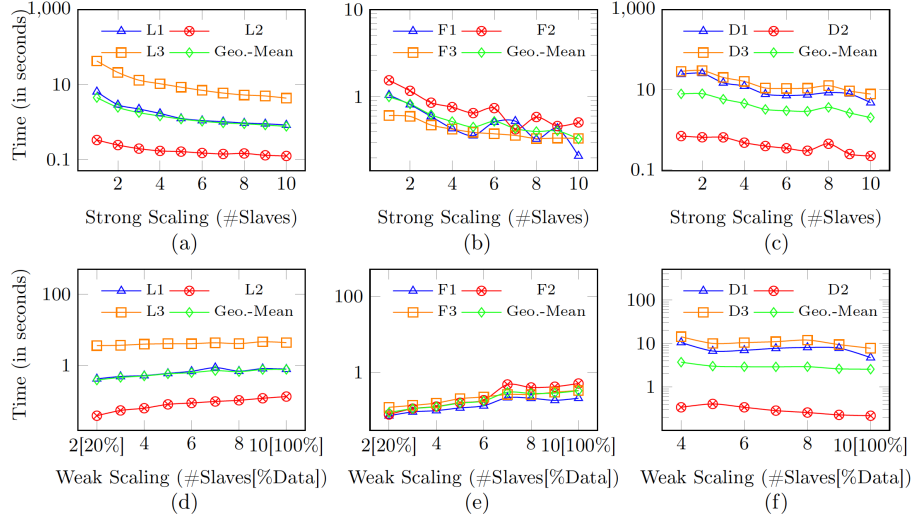
**B. Freebase-500M.** For Freebase, we considered three queries (F1, F2, F3) which we designed along the lines of the L1, L2, L3 LUBM queries. The performance of TriAD\* for Freebase-500M shows a similar behavior as the one we observed for LUBM-500M. The results are shown in Table 1(b). F1 again consists of a single, non-selective reachability join. For this query, TriAD\* performs better than Virtuoso under a *cold* cache and has comparable performance to Virtuoso in a *warm* cache setting. For the selective query F2, which comprises of regular joins and a single reachability join, TriAD\* performs better than Virtuoso in the cold-cache setting, but Virtuoso with a warm cache outperforms TriAD\* in the both centralized and distributed settings. For the non-selective query F3, which comprises of multiple reachability joins along with regular joins, TriAD\* performs significantly better than Virtuoso under both a cold and warm cache. We remark that, for F3, Virtuoso tends to report different results over repeated runs of the query, which indicates problems with their current support for property paths.

**C. DBpedia.** We once more considered three queries (D1, D2, D3) consisting of a mixture of relational joins and graph-reachability predicates for DBpedia. The runtime performance of TriAD\* in comparison with Virtuoso is shown in Table 1(c). Also here, TriAD\* continues to perform very well compared to Virtuoso under both cold and warm cache settings.

## 6.2 Scalability Tests

We finally evaluated the scalability of TriAD\* for SwPP queries by varying the number of slaves from 1 to 10. For this evaluation, we again considered

LUBM-500M, Freebase-500M and DBpedia. The results under *strong scaling* are shown in Figure 4(a) for LUBM-500M, in Figure 4(b) for Freebase-500M, and in Figure 4(c) for DBpedia, respectively. As our last series of runs, we also evaluated the performance of TriAD\* under *weak scaling*, by increasing the size (from 20%–100%) of the collections as well as the number of slaves (from 2–10) in equal proportions. The results are shown in Figure 4(d)–(f).



**Fig. 4.** Strong (a)–(c) and weak (d)–(f) scalability of SwPP queries

## 7 Conclusions

We presented the architecture of TriAD\*, which to our knowledge is the currently fastest, distributed RDF engine that explicitly tackles the processing of property paths according the recently updated SPARQL 1.1 specification. Building on top of our TriAD engine, we leverage its multi-threaded and asynchronous query processing framework to implement a new relational query operator to tackle the kind of generalized graph-pattern queries that arise in SPARQL 1.1. Our evaluation over both real-world and synthetic RDF collections confirm that TriAD\* achieves very significant gains compared to the only currently available, native RDF store that supports SPARQL 1.1 with property paths. As for future work, we intend to extend TriAD\* to support also more complex property-path variants, including paths with length restrictions, shortest paths, as well as paths with more general regular expressions.

## References

1. <http://www.w3.org/TR/rdf-schema/>.
2. <http://www.w3.org/TR/rdf-sparql-query/>.

3. <http://www.w3.org/TR/sparql11-property-paths/>.
4. P. Cudré-Mauroux et al. NoSQL databases for RDF: an empirical evaluation. In *ISWC*, pages 310–325, 2013.
5. O. Erling and I. Mikhailov. Virtuoso: RDF Support in a Native RDBMS. In *SWIM*, pages 501–519, 2009.
6. W. Fan, X. Wang, and Y. Wu. Performance guarantees for distributed reachability queries. *PVLDB*, 5(11):1304–1315, 2012.
7. D. Fensel, S. Decker, M. Erdmann, and R. Studer. Ontobroker: The Very High Idea. In *FLAIRS*, pages 131–135, 1998.
8. T. M. Forum. MPI: A Message Passing Interface, 1993.
9. S. Gao and K. Anyanwu. PrefixSolve: efficiently solving multi-source multi-destination path queries on RDF graphs by sharing suffix computations. In *WWW*, pages 423–434, 2013.
10. A. Gubichev, S. J. Bedathur, and S. Seufert. Sparqling Kleene: fast property paths in RDF-3X. In *GRADES*, pages 14:1–14:7, 2013.
11. S. Gurajada, S. Seufert, I. Miliaraki, and M. Theobald. TriAD: a distributed shared-nothing RDF engine based on asynchronous message passing. In *SIGMOD*, pages 289–300, 2014.
12. S. Gurajada, S. Seufert, I. Miliaraki, and M. Theobald. Using graph summarization for join-ahead pruning in a distributed RDF engine. In *SWIM*, pages 1–4, 2014.
13. S. Gurajada and M. Theobald. Distributed set reachability. In *SIGMOD*, pages 1247–1261, 2016.
14. S. Hagedorn, K. Hose, and K. Sattler. SPARQLing Pig - Processing Linked Data with Pig Latin. In *BTW*, pages 279–298, 2015.
15. J. Huang, D. J. Abadi, and K. Ren. Scalable SPARQL Querying of Large RDF Graphs. *PVLDB*, 4(11):1123–1134, 2011.
16. G. Karypis and V. Kumar. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, 1998.
17. T. Neumann and G. Weikum. The RDF-3X engine for scalable management of RDF data. *The VLDB Journal*, 19(1):91–113, 2010.
18. M. Przyjaciół-Zablocki, A. Schätzle, T. Hornung, and G. Lausen. RDFPath: Path query processing on large RDF graphs with MapReduce. In *ESWC*, pages 50–64, 2011.
19. K. Rohloff and R. E. Schantz. Clause-iteration with MapReduce to scalably query datagraphs in the SHARD graph-store. In *DIDC*, pages 35–44, 2011.
20. S. Sakr and G. Al-Naymat. Relational Processing of RDF Queries: A Survey. *SIGMOD Rec.*, 38(4):23–28, 2010.
21. S. Seufert, A. Anand, S. J. Bedathur, and G. Weikum. FERRARI: flexible and efficient reachability range assignment for graph indexing. In *ICDE*, pages 1009–1020, 2013.
22. L. Sidirouros, R. Goncalves, M. Kersten, N. Nes, and S. Manegold. Column-store support for RDF data management: not all swans are white. *PVLDB*, 1(2):1553–1563, 2008.
23. M. Then, M. Kaufmann, F. Chirigati, T. Hoang-Vu, K. Pham, A. Kemper, T. Neumann, and H. T. Vo. The more the merrier: Efficient multi-source graph traversal. *PVLDB*, 8(4):449–460, 2014.
24. R. S. Xin, J. E. Gonzalez, M. J. Franklin, and I. Stoica. GraphX: a resilient distributed graph system on Spark. In *GRADES*, 2013.
25. H. Yildirim, V. Chaoji, and M. J. Zaki. GRAIL: Scalable Reachability Index for Large Graphs. *PVLDB*, 3(1-2):276–284, 2010.

26. K. Zeng, J. Yang, H. Wang, B. Shao, and Z. Wang. A distributed graph engine for web scale RDF data. *PVLDB*, 6(4):265–276, 2013.
27. X. Zhang, L. Chen, Y. Tong, and M. Wang. EAGRE: Towards scalable I/O efficient SPARQL query evaluation on the cloud. In *ICDE*, pages 565–576, 2013.



## A SwPP Queries

### A.1 LUBM Queries

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix ub: <http://www.lehigh.edu/~zhp2/2004/0401/univ-bench.owl#>

L1: SELECT * WHERE { ?x rdf:type ub:ResearchGroup . ?x ub:subOrganizationOf* ?y.
    ?y rdf:type ub: University. }

L2: SELECT * WHERE { ?x rdf:type ub:FullProfessor. ?x ub:headOf ?d.
    ?d ub:subOrganizationOf* ?y. ?y rdf:type ub:University.}

L3: SELECT * WHERE { ?r1 rdf:type ub:ResearchGroup . ?r1 ub:subOrganizationOf* ?y.
    ?y rdf:type ub:University . ?r2 rdf:type ub:ResearchGroup. ?r2 ub:subOrganizationOf* ?y.}
```

### A.2 Freebase Queries

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix fb: <http://rdf.freebase.com/ns>

F1: SELECT * WHERE { ?p fb:people.person.place_of_birth ?city .
    ?city fb:location.location.containedby* ?state.
    ?country fb:location.location.contains ?state. }

F2: SELECT * WHERE { ?p fb:people.person.place_of_birth ?city .
    ?city fb:location.location.containedby* ?state. ?country fb:location.location.contains
    ?state. ?p fb:award.award_winner.awards_won ?prize.
    ?p rdf:type fb:government.us_president. }

F3: SELECT * WHERE {?p fb:award.award_winner.awards_won ?prize. ?prize rdf:type* ?z .
    ?z fb:award.award_honor.ceremony> ?c.?p fb:people.person.sibling_s* ?p1.
    ?p1 fb:award.award_winner.awards_won ?prize. }
```

### A.3 DBpedia Queries

Namespace prefixes available at: <http://de.dbpedia.org/sparql?nsdecl>

```
D1: SELECT * WHERE { ?s1 rdf:type ?s. ?s rdfs:subClassOf* ?o.
    ?o owl:equivalentClass yago-res:wordnet_medium_106254669 }

D2: SELECT * WHERE { ?s foaf:isPrimaryTopicOf wiki:North_Auburn,_California .
    ?s dbpedia-owl:isPartOf* ?c. ?x dbpedia-owl:hometown ?c. ?x foaf:isPrimaryTopicOf ?r. }

D3: SELECT * WHERE { ?s dbpprop:leaderTitle ?title. ?title rdf:type ?class.
    ?class rdfs:subClassOf* ?class2.
    ?class2 owl:equivalentClass yago-res:wordnet_abstraction_100002137 .
    ?s foaf:isPrimaryTopicOf wiki:North_Auburn,_California . ?s dbpedia-owl:isPartOf* ?c.
    ?x dbpedia-owl:hometown ?c. ?x foaf:isPrimaryTopicOf ?r. }
```